JRC Scientific and Technical Reports

State of the Art in Privacy Preserving Data Mining

Igor Nai Fovino and Marcelo Masera

EUR 23068 EN - 2008

JRC EUROPEAN COMMISSION

ipsc Institute for the Protection and Security of the Citizen

The Institute for the Protection and Security of the Citizen provides research-based, systems-oriented support to EU policies so as to protect the citizen against economic and technological risk. The Institute maintains and develops its expertise and networks in information, communication, space and engineering technologies in support of its mission. The strong cross-fertilisation between its nuclear and non-nuclear activities strengthens the expertise it can bring to the benefit of customers in both domains.

European Commission
Joint Research Centre
Institute for the Protection and Security of the Citizen

Contact information
Address: Via E Fermi I-21020 Ispra (VA) ITALY
E-mail: igor.nai@jrc.it
Tel.: +39 0332786541
Fax: +39 0332789576

http://ipsc.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/

JRC JRC42698

EUR 23068 EN
ISSN 1018-5593

Luxembourg: Office for Official Publications of the European Communities

Printed in Italy

State of the Art in Privacy Preserving Data Mining

Igor Nai Fovino, Marcelo Masera

16th January 2008

Contents

Introduction

Introduction We live today in the Information Society. Every second, millions of information are stored in some "Information Repository" located everywhere in the world. Every second, millions of information are retrieved, shared and analyzed by someone. On the basis of the information stored in a database, people develops economical strategies, makes decision having an important effect on the life of other people. Moreover, this information is used in critical applications, in order to manage and to maintain for example nuclear plants, defense sites, energy and water grids and so on. In few words, "Information is a precious asset for the life of our society". In such a scenario, the information protection assumes a relevant role.

A relevant amount of information stored in a database is related to personal data or, more in general, to information accessible only by a restricted number of users (we call this information "Sensitive Information"). The concept of *Information Privacy* is then relevant in this context.
In the scientific literature several definitions exist for privacy (we describe in depth this concept in the following chapters). At this moment, in order to introduce the context, we briefly define the privacy as a:

"Limited access to a person and to all the features related to the person".

In the database context, the information privacy is usually guaranteed by the use of access control techniques. This approach guarantees an high level of privacy protection against attacks having as final goal the direct access to the information stored in a database.

Access control methods, however, result nowadays prone to a more sophisticated family of privacy attacks based on the use of data mining techniques.

Data Mining techniques has been defined as "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [41]. In other words, by using DM techniques, it is possible to extract new and implicit information from known information.

As a general principle, the data mining technology is neutral with regard to privacy. However the goal for which it is used can be either good or malicious. Roughly speaking, even if data mining has expanded scientific (and not only) investigation possibilities by automating data reduction procedures to enable scientists to effectively exploit immense datasets, on the other hand the malicious use of such techniques is a serious threat against the privacy protection.

In a typical database, a large number of relationships (both explicit and implicit) exist between the different information. These relationships constitute a potential privacy breach. In fact, by applying some access control methods, one can avoid the direct access to sensitive information. However, a sensitive information, by the presence of these relationships, influences in some way other information. It is then possible, by applying DM techniques to the accessible information, to reconstruct indirectly the sensitive information, violating in such a way the privacy property.
These considerations have an important consequence:

Even if I protect a sensitive information using a control access method, I cannot guarantee that a malicious agent, by the use of some Data Mining technique will be able to guess the same information for which he has not the right to have access, analyzing apparently not related and accessible information.

Recently, a new class of data mining methods, known as privacy preserving data mining (PPDM) algorithms, has been developed by the research community working on security and knowledge discovery. The aim of these algorithms is the extraction of relevant knowledge from large amount of data, while protecting at the same time sensitive information. Several data mining techniques, incorporating privacy protection mechanisms, have been developed that allow one to hide sensitive itemsets or patterns, before the data mining process is executed. For example privacy preserving classification methods, prevent a miner from building a classifier which is able to predict sensitive data.

All the PPDM techniques actually presented in literature prevent data disclosure by modifying some characteristics of the database. A database can be identified as a model of the real world. Every time one modifies the data contained in a database, the world described by the database is modified. If this modification downgrades the quality of the model, its effects can be potentially dangerous. We describe in the next chapters some tests we have performed. These tests show that the impact of PPDM algorithms on the database data quality are relevant. For this reason, we believe that it is mandatory, especially for the databases containing critical data, to take in consideration this aspect.

In other words, the question can be the following: "Which is the trade-off between privacy e data quality?" or, more specifically, "Is it possible to hide sensitive information without damaging the other information contained in the database?".

The presented problem is an open question. In this report, we explore the aspects related to the database sanitization (i.e. the process by which data is modified in order to hide sensitive information), using as main criteria the Quality of the Data. In order to achieve this goal, it is necessary to understand what is data quality. DQ is a very complex concept, incorporating both objective and subjective parameters. In the context of PPDM, this implies that, in order to preserve the DQ, a sanitization algorithm needs to understand the meaning of the information, or, more simply, it needs to know what is relevant, and then which information must be preserved, in the particular context of the target

database. More specifically, in order to preserve the data quality during the sanitization phase, we introduce a schema allowing to represent relevance and meaning of the information contained in a target database and the relationships between these data. Moreover, we present a new data hiding approach based on the use of this type of information.

As explained before, the privacy problem is today very relevant. The PPDM techniques represent actually the "State of the Art" against the privacy attacks in the database context. Due to their particular nature, however, every PPDM technique gives good results only under certain constraints. Such constraints are mainly related to the database structure, the type of data, the type of information mined and so on. For this reason, different PPDM techniques guarantee a different level of "Privacy Service" on the same database. Therefore, in a context in which the people responsible for the sensitive information contained in a database must, by low [1], guarantee the highest level of security against the risk of privacy breaches, it is important to have a methodology allowing one to identify the most suitable PPDM algorithm for a target database.

In the scientific literature, a common and generally accepted set of evaluation parameters does not exist. There are in fact some open questions:

- Which set of evaluation criteria must be considered?
- Which one must be considered the most relevant in a set of evaluation criteria?
- How can we compare algorithms based on completely different approaches?

In this work we present our results in the identification of a set of general evaluation parameters for privacy preserving data mining algorithms. Moreover, in this context we introduce some new concepts like the Privacy Level. We present the Information Schema, allowing to represent the meaning and the relevance of the information contained in a database and we show how to use this schema in assessing the sanitization impact on the database data quality. Finally we present a complete three-step PPDM Evaluation Framework.

[1] Several national government recently have introduced new laws in order guarantee the citizen's right of privacy

Chapter 1

Data Mining

Data Mining can be assumed as a particular type of Knowledge discovery process. It can be defined as the analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways, understandable and useful to the owner.

Historically, data mining can be viewed as a result of the natural evolution of information technology. In fact, an evolutionary path has been witnessed in the database industry in the development of the following functionalities:

- Data collection and database creation.
- Data management.
- Data analysis and understanding.

Figure 1.1 shows such an evolution. As it is possible to see, starting from the initial principle of data collection, the evolution of database systems is completely oriented to enforce the data manipulation and representation capabilities. The last frontier in this type of evolution is represented by the possibility to extract from simple data, knowledge not immediately evident.

1.1 A definition of Knowledge

The concept of information, or better, the concept of'*knowledge*, is the basis of the work presented in this work. It is thus necessary to define this important concept. Several definitions of *Knowledge* have been proposed. However, in order to introduce at least such a definition, it is necessary to introduce before the bricks by which knowledge is built:

- **Data**: it is a set of discrete, objective facts about events. In an organizational context, data is most usefully described as structured records of transactions. When a customer at the supermarket buys fish and onions,

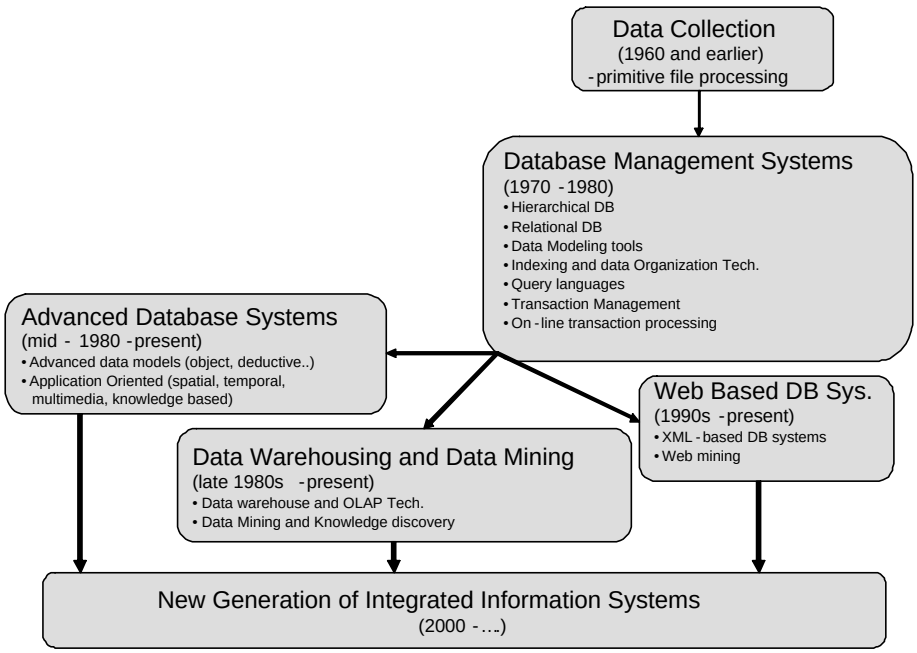

Figure 1.1: The evolution of database technology

this transaction can be partly described by data: when he made the purchase, how many onions he bought, etc. The data tells nothing about why the customer bought onion with fish or way the customer paid using a credit card or a debit card. P. Drucker [29] said that information is "data endowed with relevance and purpose," which of course suggests that data by itself has little relevance or purpose.

- **Information**: in the scientific literature, some people equates *information* with meaning [65]. Hearing a statement is not enough to make an event an informative act; its meaning must be perceived to make the statement informative. Arguing against this approach, Bar Hillel points out that "it is psychologically almost impossible not to make the shift from the one sense of information, ... i.e. information = signal sequence, to the other sense, information = what is expressed by the signal sequences" [9]. In another approach, information is often understood in terms of knowledge that is transmitted to a sentient being. For example, information may be understood as "that which occurs within the mind upon the absorption of a message" [75]. Another useful analogy is with the structure of systems that is viewed by some as being equivalent to information. Thus, "information is what remains after one abstracts from the material aspects of physical reality" [80]. However, a good general definition of information is given by Losee in [61]: *information is produced by all processes*

(in every context)and the values associated to the processes output is the information.

- **Knowledge**: the difference between knowledge and information, may be extremely thin. Epistemologists spend their lives trying to understand what it means to know something; obviously, it is not the scope of this chapter to cover such question. In any case, we can think of knowledge as a set of information explicitly and implicitly correlated. Knowledge is a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information. From a formal point of view, Frawley, Shapiro and Matheus in [41] give an interesting definition of Knowledge:

  Definition 1 *Let F a set of facts, L a language, and C a measure of certainty, a pattern is a statement $S \in L$ that describes relationships among a subset FS of F with a certainty c, such that S is simpler (in some sense) than the enumeration of all facts in FS. A pattern that is interesting and certain enough is called knowledge.*

Referring to Definition 1, it is now possible to provide a definition of the notion of knowledge discovery process.

Definition 2 *A process that monitors the set of facts described in a database and produces patterns according to Definition 1 is a* ***Knowledge discovery process****.*

As we will see in the next section, the data mining process, takes a set of data contained in a database and deduces a set of patterns similar to the patterns described in the *Definition 1*. Therefore, we can deduce that the data mining process is a particular type of knowledge discovery process.

1.2 Data Mining Concepts

From a logical point of view the Data Mining process generally provides four main logical functionalities:

- **Data Characterization:** Summarization of the general characteristics or features of a target class of data.
- **Data Discrimination:** Comparison of the general features of target class of data with respect to the general features of a contrasting class.
- **Association Analysis:** Identification of association rules showing certain attribute value condition.
- **Classification:** Identification of a set of data models that describe and distinguish data classes and concepts.

These functionalities even correspond to the type of patterns that can be mined from a database. It is not fully correct to speak of Data Mining as a homogeneous field of research. Data Mining can be assumed to be a combination of techniques and theories from other research fields (Machine Learning, Statistic, database systems). For this reason, different classification schemes can be used to categorize data mining methods. As explained in [17], it is possible to classify Data Mining techniques by adopting different metrics:

- **What kinds of target databases:** relational databases, transaction databases, object-oriented databases, deductive databases, spatial databases, temporal databases, multimedia databases, heterogeneous databases, active databases, legacy databases.
- **What kind of knowledge to be mined:** as we have mentioned, different kinds of knowledge (or patterns) exist that it is possible to extract from a database: association rules, classification rules, clusters.
- **What kind of techniques to be utilized:** Data mining algorithms can also be categorized according to the driven method into autonomous knowledge miner, data-driven miner, query-driven miner and interactive data miner.

In this brief overview of data mining techniques we will follow the "Kind of Knowledge" classification.

1.3 Association Rule Mining

The main goal of Association Rule Mining is to discover frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories. An association rule is an expression $X \rightarrow Y$, where X and Y are sets of items. The idea of mining association rules originates from the analysis of market-basket data where rules like "A customer that buys products x_1 and x_2 will also buys product y with probability c%" can be found. More formally:

Definition 3 *Let $J = \{i_1, i_2, i_3, i_n\}$ be a set of items. Let D a set of database transactions where each transaction T is a set of items such that $T \subseteq J$. Let A be a set of items. A transaction T contains an itemset A if and only if $A \subset T$. An **association rule** is an implication of the form $A \rightarrow B$ where $A \subset J$, $B \subset J$ and $A \cap B = \phi$.*

Definition 4 *The rule $A \rightarrow B$ holds in the transaction set D with a **support** s where s is the percentage of transactions in D that contains $A \cup B$.*

Definition 5 *The rule $A \rightarrow B$ has a **confidence** c in the transaction set D if c is the percentage of transaction in D containing A that also contain B*

From a mathematical point of view, the support and the confidence can be assumed equal to:

$$Supp(A \rightarrow B) = \frac{|A \cup B|}{|D|} \tag{1.1}$$

$$Conf(A \rightarrow B) = \frac{supp(A \cup B)}{supp(A)} \tag{1.2}$$

TID	**Items**
T100	I_1, I_2, I_5, I_9
T101	I_4, I_6
T102	I_3, I_4, I_8
T103	I_1
T104	I_1, I_3, I_9

Table 1.1: Example of Transactional Database

Roughly speaking, confidence denotes the strength of implication and support indicates the frequencies of the occurring patterns in the rule. Usually, especially in sparse database, it is possible to extract a wide number of rules. Intuitively not all these rules can be identified as "Of Interest". Piaterstky-Shapiro [74] and Agrawal [3] therefore introduced the concept of *strong rule*, that is, a rule with a reasonable (under such criteria) level of support and confidence. Traditionally, the association rule mining task consists in identifying a set of strong association rules. Srikant and Agrawal in [93], explore and expand the idea of *interesting rule*. More in deep, Agrawal et al. [3] and Park et al. [72] decompose the challenge of Association Mining into two main steps:

1. Identify the large itemsets, that is the sets of itemsets with transaction support above a apriori determined minimum support s

2. Use the large itemsets to magnify the association rules contained in the database

As it is evident, the more time consuming task is the first one. In fact, a linearly growing number of items implies an exponential growing number of itemsets that need to be considered. Figure 1.2 shows an example in which we have I=1,2,3,4 where I is the set of items.

The frequent itemsets are located on the top of the figure whereas the infrequent ones are located in the lower part. The bold line separates the frequent itemsets (that will constitute the strong rules) from the not frequent itemsets. Usually, algorithms that execute the previously described *step one*, try to identify this border line in order to reduce the space of possible itemsets. The two main exploration strategies used to identify this border (breadth-first search (BFS) or depth-first search (DFS)) are used to classify the Association rule

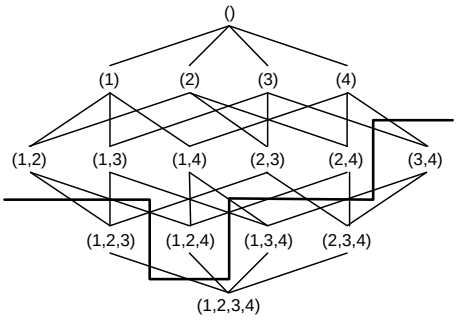

Figure 1.2: A Lattice for I=1,2,3,4

Mining Algorithms with respect to the strategy used for the support evaluation. In detail, there are two main strategies: a common and simple approach to determine the support value of an itemset is to directly count its occurrences in the database. Another approach is to determine the support values of candidates by set intersections. In this case, assuming that each transaction is uniquely identified by a code TID (Transaction Identifier), it is possible to associate, to each item, the list of the TID (TIDlist) of the transactions containing the item. Assuming to have an itemset $I = (a, b)$, the support is equal to

$$Sup = \frac{|TIDlist(a) \cap TIDlist(b)|}{|database|} \tag{1.3}$$

Figure 1.3 shows a simple classification of the algorithms for association rule mining. In what follows we give a brief description of these methodologies.

- **BFS/Counting:** The Apriori [5] algorithm is a good representative of this category of algorithms. Apriori counts all candidates of a cardinality k together in one scan over the database. The most important operation is looking up for the candidates in every transactions. In order to perform this task, in [5] Agrawal Srikant introduced an *hashtree structure*. The items in each transaction are used to descend in the hashtree. Every time a leaf is found, a set of candidates if found. Then these candidates are searched in the transaction that has been encoded as a bitmap before. In the case of success, the counter of the candidate in the tree is incremented. DIC (Dynamic Itemset Counting) [14] is a variation of Apriori Algorithm that softens the strict separation between counting and generating candidates, trying to optimize such operations.

- **BFS/Intersecting:** The Partition Algorithm [85] is a derivation of the Apriori algorithm that uses set intersections to determine support values. The Partition Algorithm divides the database into several mini-database that are independently analyzed. After determining the frequent itemsets

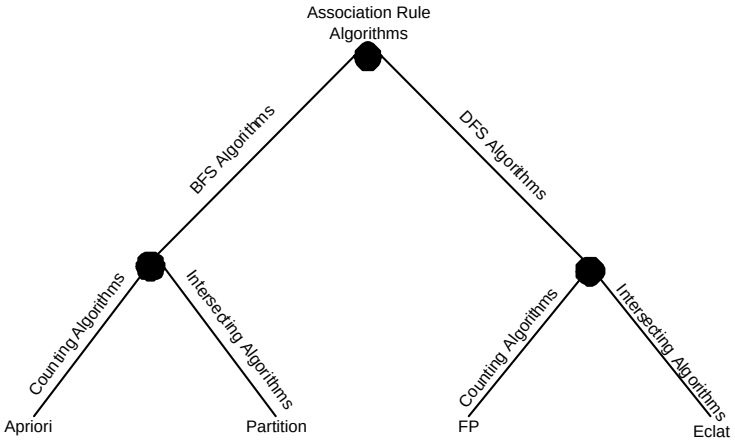

Figure 1.3: An association rule algorithm classification

for each mini-databases, a global scan is made to identify the globally frequent itemsets.

- **DFS/Counting:** Under DFS the candidate sets consist only of the itemsets of a single node as the ones described in the lattice in Figure 1.2. That implies that a scan on the database will be executed for each node. This is obviously a very time consuming operation. In order to solve this problem, the FP-growth algorithm was proposed in [47]. In a pre-analysis phase, the FP-growth algorithm builds highly condensed representation of the transaction data called FP-tree. This new tree is then used to count the support of the interesting itemset.

- **DFS/Intersecting:** such an approach, combines the DFS with the intersecting methodology. Eclat [108] keeps the Tidlists (described before) on the path from the root down to the class currently investigated in memory. The partition algorithm is then applied to these Tidlists.

1.4 Clustering Techniques

The main goal of a clustering operator, as showed in Figure 1.4, is to find a reasonable segmentation of the records (data) according to some criteria [46]. Each segment (cluster), consists of objects that are similar among themselves and dissimilar from objects of other groups. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters can be assumed as a unsupervised learning and the resulting system represents a data concept.
It is possible to give here a very short categorization of clustering algorithms:

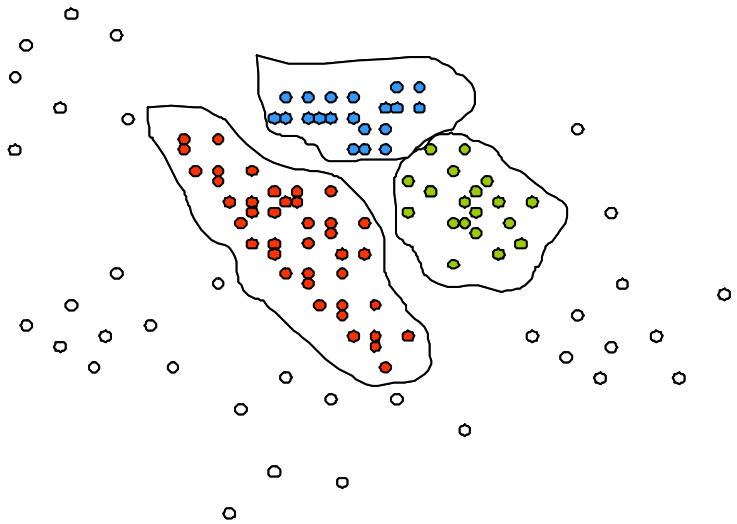

Figure 1.4: An example of cluster grouping

- **Hierarchical Methods:**Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendrogram. In particular such an algorithm initializes a cluster system as a set of singleton clusters (agglomerative case) or a single cluster of all points (divisive case) and proceeds iteratively with merging or splitting of the most appropriate cluster(s). Some interesting algorithms based on such approach are the SLINK algorithm [90], the CLINK algorithm [22] and the AGlomerative NESting (AGNES) algorithm [55]. These algorithms can be identified as Linkage metrics-based hierarchical clustering methods. More recently, Guha et al. [44] introduced the hierarchical agglomerative clustering algorithm CURE (Clustering Using REpresentatives). This algorithm can be identified as a good example of the "Hierarchical Clusters of Arbitrary Shapes" clustering subclass. In such a context, the CHAMELEON algorithm [54] introduce the use of dynamic modeling in cluster aggregation

- **Partitioning Relocation Methods:** they divide data into several subsets. Because checking all possible subsets is computationally infeasible, some greedy heuristics are used in order to obtain an iterative optimization. This introduce the concept of *Relocation Schema*, that reassigns the points in the different clusters during this process. Some of these methods can be classified as probabilistic models [63]. In such a context, the SNOB algorithm [101] and the AUTOCLASS algorithm [16] have a significant position in the scientific literature. The algorithms PAM and CLARA [55] are two examples of the subclass of *k-medoids methods* in which a cluster is represented by one of its points. Finally, the most well known algorithm

in this class is the k-means algorithm [49]. Such an algorithm represent each of k clusters by the mean (or weighted average) c of its points, the so-called centroid.

- **Density-based Partitioning:** such a class of algorithms represents the implementation of the idea that *an open set in a Euclidean space can be divided into a set of its connected components*. The discriminant is then the density; in this way, a cluster grows in any direction allowing to have clusters with not well defined or pre-fixed shape. The algorithm DBSCAN (Density Based Spatial Clustering of Applications with Noise) [35] and the DENCLUE (DENsity-based CLUstEring) algorithm [50] represent the reference point in this context.

- **Methods Based on Co-Occurrence of Categorical Data:** these methods are developed in order to identify clusters in the context of categorical database. We recall here that a categorical database is one in which the values of the items can assume a limited range of fixed values. A typical example is the Market Basket database, in which each record contains sequences of 0 and 1 representing what the customer has in his basket. In this context cluster must be created searching for the co-occurrence between the different records. The Rock algorithm [45] and more recently the SNN (Shared Nearest Neighbors) algorithm [34] are representative of this class of methods

1.5 Classification Techniques

Data classification is the process which finds the common properties among a set of objects in a database and classifies them into different classes, according to a classification model. As showed in Figure 1.5, the basic idea of Classification Techniques is to use some limited set of records, named *Sample* or *Training set*, in which every object has the same number of items of the real database, and in which every object has already associated a label identifying its classification. The objective of the classification methodologies can be summarized as follow:

- **Sample set analysis**: the sample set is analyzed in order to produce a description or a model for each class using the features available in the data.

- **Accuracy Evaluation:** the accuracy of the model is evaluated. Only if the accuracy is over a certain threshold, the model will be used in the following step.

- **Test data Analysis:** using the model previously obtained a classification of new test data is executed. Moreover the model can be used to improve an already existing data description

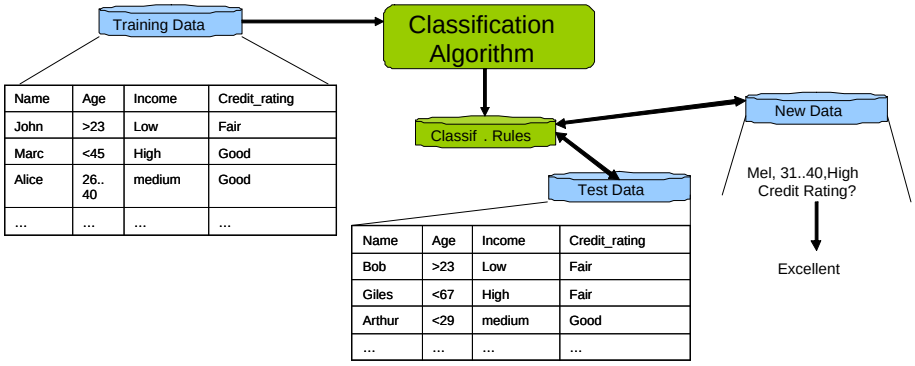

Figure 1.5: An example of classification

The possible approaches to classification, are, traditionally divided into four main classes: *Decision Tree Induction, Bayesian Classification, Backpropagation and Classification based on Association Rule Mining.* In what follows, we give a brief description of these techniques.

- **Decision Three Induction:** It is a supervised learning method that constructs decision trees from a set of examples [76]. ID-3 [76] is a good example of Decision Three Induction based method for categorical data. It adopts an information-theoretic approach aimed at minimizing the expected number of tests to classify an object. C4.5 [77] is an evolution of ID-3 that allows classification even on numerical data. Another example of such type of algorithms is PUBLIC [78]. From the evaluation point of view, several approaches exist. For example, ID-3 adopts the following function:

$$i = \sum (p_i ln(p_i)) \tag{1.4}$$

  where p_i represents the probability that an object is in class i. Another example is the *gini index* [13] measured as:

$$Gini(T) = \sum 1 - p_i^2 \tag{1.5}$$

  where p_i is the relative frequency of class i in T. In Decision tree methodologies the *pruning strategy* is fundamental in order to eliminate anomalies due to noise or outliers. Some examples of decision tree pruning algorithms are in Mehta, Agrawal and Rissanen [64] and Rastogi and Shim [79].

- **Bayesian Classification**: Bayesian decision theory is the basis of these classification methods. Consider the following simple example: Let CP be a group classification problem, in which each object has an associated attribute vector x of dimensions z. We say that if the object is member of a group, a membership variable w, takes as value w_j. If we define $P(w_j)$

as the prior probability of group j and $f(x|w_j)$ the probability density function, according to the Bayes rule :

$$P(w_j|x) = \frac{f(x|w_j)P(w_j)}{f(x)} \tag{1.6}$$

If the size of the group is K, the density function is then equal to:

$$\sum_{j=K}^{j=1} f(x|w_j)P(w_j) \tag{1.7}$$

From this equation it is possible to derive the classification error probability if we choose w_j, that is equal to:

$$P(error|X) = 1 - P(w_j|X) \tag{1.8}$$

Such a definition is on the basis of the Bayes classification methods [30]. However, the world of bayesian classificators is extremely diversified and it is not possible to present here all the algorithms developed nor a complete survey on the state of the art (we recall here that this report work has as target PPDM algorithms and not Data Mining algorithms). In any case, for the sake of completeness we report here the most interesting approaches. Domingos and Pazzani [28] have presented an interesting study on the power of naive baysian classifier. Algorithms using belief networks are presented by Russel, Binder, Koller and Kanazawa [82]. Also the approach of Lauritzen [57]is very interesting. Such an approach is based on the use of learning belief network with hidden variables

- **Backpropagation**: the backpropagation is a neural network algorithm. It performs learning on a multilayer feed-forward network layer. The inputs of this network are the attributes measured for a training set. Roughly speaking the backpropagation technique works as follow:

  1. It Initializes the network weights.
  2. It takes as input the sample set.
  3. It makes a prediction.
  4. It compares prediction with the actual class.
  5. It modify iteratively the weight in order to obtain a better prediction.

  The backpropagation algorithm was presented originally by Rumelhart et al. [83]. Starting from this algorithm some variations has been proposed, introducing for example *alternative error functions* [48], dynamic adjustment of the network topology [38].

Chapter 2

Privacy Preserving Data Mining

The rapid evolution of Data Mining has given a lot of benefits in data analysis and Knowledge discovery. Such technology has been applied to a wide range of fields from financial data analysis [26] to intrusion detection systems [58,107] and so on. However, as every technology, it is not good or malicious by definition. That is obviously even for Data Mining. Consider a database that has to be shared among several users. Some data contained in the database are protected using access control methods in order to guarantee that only authorized people are allowed to have access to these sensible information. The use of data mining techniques by people with a limited access to the database can easily circumvent the access control system. This type of "attack" cannot easily be detected, because, usually, the data used to guess the protected information, is freely accessible. To address this problem is the main goal of a relatively new field of research named Privacy Preserving Data Mining

2.1 The problem of "Preserving Privacy"

As described in [60], to define correctly the privacy preserving problem, it is necessary first define a data representation. One of the most popular data representation is by data tables [73]. Adopting such a representation, it is easy to think about a relational database as a collection of data tables linked by some "relations". Starting from the data representation suggested by Pawlak [73], a data table can be defined as follow:
Definition 6 *Data Table*
A data table is a pair T = (U,A) such that

- U is a nonempty finite set, called the universe
- A is a nonempty finite set of primitive attributes

- Every primitive attribute $a \in A$ is a total function $a : U \rightarrow V_a$, where V_a is the set of values of a, called the domain of a.

The attributes associated with a data table can be generally divided into three sets [51]:

- **Key Attributes:** They are used to identify a data record. They directly identify individuals (directly associated to elements in the universe U), so we can assume that in a big number of situations they are always masked off in response to queries.
- **Public attributes:** Their values are normally accessible to the authorized people. These attributes, if not appropriately generalized or protected, may be used to break the privacy of an individual record.
- **Confidential attributes:** The values that are considered sensible and that we want to absolutely protect.

Wang et. all, in [51], in order to well identify the privacy problem, suggest to reorganize a data table as a data matrix mapping the universe U to the rows of the matrix taking into account the previously given attributes classification. More specifically, a data matrix T is a $n \times m$ matrix $[t_{ij}]_{n \times m}$ such that for each $1 \leq j \leq m$, $t_{ij} \in V_j$, where V_j is the domain of attribute j. We can assume now that the attributes are ordered in the matrix T in such a way the attributes at positions $1..m_1$ are public attributes and the attributes at positions $m_1 + 1..m_2$ are the confidential ones. In such a way, we can identify two sub-matrix *pub(T)* and *conf(T)* containing the two different type of attributes. Ideally, as introduced in [51], database manager or people with full access rights are in possession of a triple (U,T,J) where U is the universe of individuals, T is the data matrix and J is a function $J : U \rightarrow t_1...t_n$ that assigns to each individual a data record. On the other hand, a user who accesses the database, also has another triple (U,pub(T),J'), where J' is a function defined as $J' : U \rightarrow pub(t_1), pub(t_2)..., pub(t_n)$. Given this formal definition, it is possible now to characterize the Privacy Preserving Problem as follow:

How can T be modified in such a way the user would not know any individual's confidential attribute values and the modified matrix is kept as informative as possible?

Mapping this problem on the specific case of *Data Mining*, it can be rewritten as follows:

How can T be modified in such a way the use of data mining techniques do not give as result any individual's confidential attribute or confidential "information", while preserving at the same time the utility of such data for authorized uses?

The previous characterization introduces some interesting and open issues that represent the core of PPDM research field:

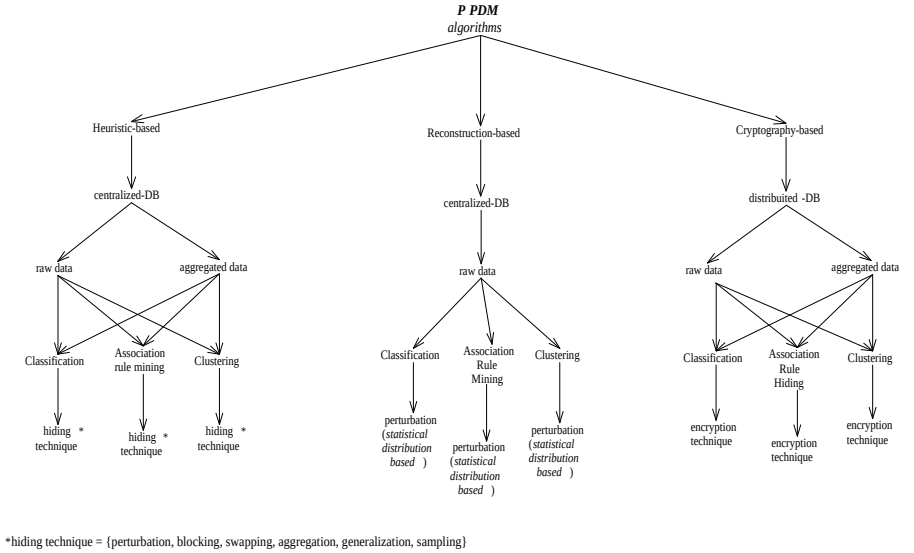

Figure 2.1: A taxonomy of the developed PPDM algorithms

- Which are the appropriate modification techniques to adopt in order to protect data?
- What is the meaning of "Data Utility"?
- How to evaluate the quality of such methods?

Some of these questions will be the topic of this report. In the next part of this section we introduce some answers to the first sentence by giving a brief overview of the different methods actually used to protect data by malicious data mining. In the next chapters we give, by the definition of a new data quality approach to the PPDM problem, some answers to the other two questions.

2.2 Methods Taxonomy

Several methods have been developed to solve the PPDM problem. These methods are usually based on very different concepts and technologies and have been originally developed to solve different aspects of the PPDM problem. In this section, we try to give a first complete taxonomy and classification of the PPDM algorithms, based on the analysis by Verykios et al. [99].

In such analysis various PPDM techniques are classified according to five different dimensions:

1. Data distribution (centralized or distributed).

2. The modification applied to the data (encryption, perturbation, generalization, and so on) in order to sanitize them.

3. The data mining algorithm which the privacy preservation technique is designed for.

4. The data type (single data items or complex data correlations) that needs to be protected from disclosure.

5. The approach adopted for preserving privacy (heuristic, reconstruction or cryptography-based approaches).

The first dimension refers to the distribution of data. Some of the approaches have been developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios can also be classified as horizontal data distribution and vertical data distribution. Horizontal distribution refers to the cases where different database records reside in different sites (as in a distributed system), while vertical data distribution, refers to the cases in which all the values for different attributes reside at different sites. The second dimension refers to the data modification scheme. In general, data modification is used in order to modify the original values of a database that needs to be released. Possible modifications include:

- Perturbation which is accomplished by altering an attribute value by assigning it a new value (i.e., changing a 1-value to a 0-value, or adding noise).

- Blocking, which is the replacement of an existing attribute value with a null value denoted usually by "?".

- Aggregation or merging, which is the combination of several values into a coarser category.

- Swapping, which refers to interchanging values of individual records.

- Sampling, which refers to releasing only a sample of the data.

The third dimension refers to the data mining algorithm, for which the data modification is taking place. This is actually something that is not known beforehand, but it facilitates the analysis and design of the data hiding algorithm. We have included the problem of hiding data for a combination of data mining algorithms into our future research agenda. For the time being, various data mining algorithms have been considered in isolation of each other. Among them, the most important ideas have been developed for classification data mining algorithms, like decision tree inducers, association rule mining algorithms, clustering algorithms, rough sets, and Bayesian networks.

The fourth dimension refers to whether raw data or aggregated data should be hidden. The complexity of hiding aggregated data in the form of rules is of course higher, and for this reason, heuristics have been developed.

The last dimension, which is the most important, refers to the privacy preservation technique used for the selective modification of the data. Selective modification is required in order to achieve higher utility for the modified data given that the privacy is not jeopardized. The techniques that have been developed can be classified into:

- Heuristic-based techniques like adaptive modification that modifies only selected values with the goal of minimizing the utility loss rather than all available values.
- Cryptography-based techniques, like secure multiparty computation, under which a computation is secure if at the end of the computation, no party knows anything except its own input and the results.
- Reconstruction-based techniques where the original distribution of the data is reconstructed from the randomized data.

Figure 2.1 shows a taxonomy of the existing PPDM algorithms according to those dimensions. Obviously, it represents a first organization in this new area and does not cover all the possible PPDM algorithms. However, it provides one an overview of the algorithms that have been proposed so far, focusing on their main features. While heuristics and reconstruction-based techniques are mainly conceived for centralized datasets, cryptography based algorithms are designed for protecting privacy in a distributed scenario. Reconstruction-based algorithms recently proposed aim at hiding sensitive raw data by applying perturbation techniques based on probability distributions. Moreover, several heuristic-based approaches for hiding both raw and aggregated data through a hiding technique (perturbation, blocking, data swapping, aggregation, generalization and sampling) have been developed, first, in the context of association rule mining and classification and, more recently, for clustering techniques. We now briefly describe some of the algorithms proposed in the PPDM area.

Atallah et al. [7], propose an heuristic for the modification of the data based on data perturbation. More specifically the procedure was to change a selected set of 1-values to 0-values, so that the support of sensitive rules is lowered in such a way that the utility of the released database is kept to some maximum value. A subsequent approach reported in [21] extends the sanitization of sensitive large itemsets to the sanitization of sensitive rules. The technique adopted in this approach is either to prevent the sensitive rules from being generated by hiding the frequent itemsets from which they are derived, or to reduce the confidence of the sensitive rules by decreasing it below a user-specified threshold. These two approaches led to the generation of three strategies for hiding sensitive rules. The important aspect to mention with respect to these three strategies is the possibility for both a 1-value in the binary database turned into a 0-value and a 0-value turned into a 1-value. This flexibility in data modification has the side-effect that, apart from non-sensitive association rules that were becoming hidden, a non-frequent rule could become a frequent one. We refer to these rules as ghost rules. Given that sensitive rules are hidden, both non-sensitive rules

which were hidden and non-frequent rules that became frequent (ghost rules) count towards the reduced utility of the released database. For this reason, the heuristics in this approach must be more sensitive to the utility issues, given that the security is not compromised. A complete approach which was based on this idea, can be found in [100].

Oliveira and Zaiane [68] propose a heuristic-based framework for preserving privacy in mining frequent itemsets. They focus on hiding a set of frequent patterns, containing highly sensitive knowledge. They propose a set of sanitized algorithms, that only remove information from a transactional database, also known in the *Statistical Diclosure Control* area as *non-perturbative* algorithms, unlike those algorithms, that modify the existing information by inserting noise into the data, referred to as *perturbative* algorithms. The algorithms proposed by Oliveira and Zaiane rely on a item-restriction approach, in order to avoid the addition of noise to the data and limit the removal of real data. In the evaluation of the proposed algorithms they introduce some measures quantifying the effectiveness and the efficiency of their algorithms.

In [92], instead, Sweeney proposes a heuristic-based approach for protecting raw data through generalization and suppression techniques. The methods she proposes provide *K-Anonymity*. Roughly speaking a database is K-anonymous with respect to some attributes if there exist at least k transactions in the database for each combination of the attribute values. A database A can be converted into a new database A^1 that guarantees the K-Anonymity property for a sensible attribute by performing some generalizations on the values of the target attributes. As result, such attributes are susceptible to *cell distortion* due to the different level of generalization applied in order to achieve K-Anonymity. Sweeney measures the cell distortion as the ratio of the domain of the attribute to the height of the attribute generalization which is a hierarchy. In the same article the concept of *precision* is also introduced. Given a table T, the *precision* represents the information loss incurred by the conversion process from a table T to a K-Anonymous Table T^k. More in detail the *precision* of a table T^k is measured as one minus the sum of all cell distortions, normalized by the total number of cells. In some way, the *precision* is a measure of the data quality or the data utility of the released table, specifically conceived for PPDM algorithms adopting generalization techniques for hiding sensitive information.

A reconstruction-based technique is proposed by Agrawal and Srikant [4] for estimating the probability distribution of original numeric data values, in order to build a decision tree classifier from perturbed training data. More in detail, the question they addressed was whether, given a large number of user who want to make this perturbation, it is still possible to construct a sufficiently accurate predictive model. They suggest two algorithms for the case of classification. The algorithms were based on a Baesyan procedure for correcting perturbed distribution. This approach obviously preserve the individual privacy, in fact, reconstructing the distribution do not release any type of information related to a target individuals. They propose some measures in order to evaluate the privacy introduced by the application of these algorithms that will be presented in Chapter 4.

The reconstruction-based approach proposed by Agrawal and Aggarwal [2] is based on an Expectation Maximization (EM) algorithm for distribution reconstruction, which converges to the maximum likelihood estimate of the original distribution on the perturbed data. The basic idea of this class of algorithm is the following: by perturbing the data and reconstructing distributions at an aggregate level in order to perform the mining it is possible to retain privacy while accessing the information implicit in the original attributes. However, the problem of this technique is related with the reconstruction of the data. In fact, depending on the approach adopted, the data reconstruction may cause an information loss. Even if in some situation this information loss can be ignored, it is important to pay attention to the reconstruction process. Agrawal and Aggarwal propose the use of the EM algorithm to make the reconstruction in order to mitigate this problem

Evfimievski et al. [37] propose a framework for mining association rules from transactions consisting of categorical items, where the data has been randomized to preserve privacy of individual transactions, while ensuring at the same time that only true associations are mined. They also provide a formal definition of privacy breaches and a class of randomization operators that are much more effective in limiting breaches than uniform randomization. According to Definition 4 from [37], an itemset A results in a privacy breach of level ρ if the probability that an item in A belongs to a non randomized transaction, given that A is included in a randomized transaction, is greater or equal ρ. In some scenarios, being confident that an item be not present in the original transaction may also be considered a privacy breach. In order to evaluate the privacy breaches, Evfimievski et al. introduce some metrics that will be presented in Chapter 4.

Another reconstruction-based technique is proposed by Rivzi and Haritsa [81]. They propose a distortion method to pre-process the data before executing the mining process. Their goal is to ensure privacy at the level of individual entries in each customer tuple.

A cryptography-based technique is proposed by Kantarcioglu and Clifton [53]. They specifically address the problem of secure mining of association rules over horizontally partitioned data, using cryptographic techniques to minimize the information shared. Their solution is based on the assumption that each party first encrypts its own itemsets using commutative encryption, then the already encrypted itemsets of every other party. Later on, an initiating party transmits its frequency count, plus a random value, to its neighbor, which adds its frequency count and passes it on to other parties. Finally, a secure comparison takes place between the final and initiating parties to determine if the final result is greater than the threshold plus the random value.

Another cryptography-based approach is described in [98]. Such approach addresses the problem of association rule mining in vertically partitioned data. In other words, its aim is to determine the item frequency when transactions are split across different sites, without revealing the contents of individual transactions. A security and communication analysis is also presented. In particular, the security of the protocol for computing the scalar product is analyzed. The

total communication cost depends on the number of candidate itemsets and can best be expressed as a constant multiple of the I/O cost of the apriori algorithm.

Recently, the problem of privacy preservation in data mining has been also addressed in the context of clustering techniques. Oliveira and Zaiane [69] have introduced a family of geometric data transformation methods for performing a clustering analysis while ensuring at the same time privacy preservation. Conventional perturbation methods proposed in the context of statistical databases do not apply well to data clustering leading to very different results in clustering analysis. Therefore, they adopt some techniques proposed for image processing in order to distort data before the mining process. More in detail, they consider the case in which confidential numerical attributes are distorted in order to meet privacy protection in clustering analysis, notably on partition-based and hierarchical methods. In this specific situation, they introduce a particular transformation (GDTM), in which the inputs are a vector V composed of confidential numerical attributes and a vector N representing the uniform noise, while the output is the transformed vector subspace V_O. In their work, Oliveira and Zaiane provide some measures (see chapter 4) proving the effectiveness of their algorithm.

Chapter 3

Statistical Disclosure Control

The Statistical Disclosure Control is a discipline that seeks to modify statistical data so that they can be published without giving any information on the individual owner of these data. A Statistical Database system is a DB system that releases to the users only aggregate statistics for a subset of the entities stored in the database (for a detailed description of such type of database see for example [42, 43]). It is the policy of the system as set by the DBA that determines the criterion for defining confidential information [23]. As explained even in the context of PPDM, threats to data security are related to the risk that some previously unknown confidential data about a given entity be disclosed. A disclosure (either partial or complete) occurs [1] if through the answer to one or more queries a snooper is able to infer the exact value of a confidential information (in this case we refer to "complete" disclosure) or is able to guess a more accurate estimation of the real confidential value ("partial" disclusure)

Shoshani [89] categorizes existing statistical database confidentiality preservation strategies into five classes:

1. Limiting the response set size: according to this strategy, a statistical database refuses to answer queries when the response set size is too small. A well known problem related to this approach is that it can be attacked adopting a strategy referred to as *tracker* [24].

2. Limiting the intersection of response sets: it requires a query logging facility in order to identify intersecting queries. For many application this solution is not really feasible. In fact, to make log analysis on the fly every time a new query is issued is resource consuming.

3. Random sample queries: the DBMS calculates the result on the basis of a random subset of the response set. In order to avoid some filtering operation, it is necessary to log the subset composition in order to give the same result for the same query [24].

4. Partitioning the database: such an approach is based on the idea to cluster the different entities stored, in a number of mutually exclusive subsets, usually called *atomic population*. In such a way, in the case in which clusters are populated by more than one individual, it is possible to obtain a good level of security against some attacks. Schlorer [86] has showed that such solution in real database has some problems due to the presence of a big number of single unit clusters. In [19] Chin and Ozsoyoglu propose a solution consisting in the addition of dummy entities to the database. Even if this solution solves the problem of single unit cluster, there is the open question related on the data quality impact of this method.

5. Perturbation of data values: this strategy consists in modifying the values stored by adding some noise. In such a way, even if some data are modified, the aggregate statistics maintain their significance for a sufficiently sized response set. As showed in figure 3.1 this distortion can be applied to the entire database (like the sanitization process of PPDM) or on the fly during query execution(in the second case however issues arise related to the performance and correlation between new and past query introducing then the problem of *incremental knowledge*)

Palley and Simonoff [71] add to these strategies an additional one based on multidimensional transformation, known as *data swapping*. This strategy switches subsets of attributes between randomly selected pairs of records. In [1], Adam and Wortman identify other two approaches they define as "Conceptual". Recalling the work of Chin and Ozsoyoglu [18] and the work of Denning and Schlorer [23] these two approaches can be summarized as follow:

- Conceptual modeling approach: no operations are allowed to combine and intersect populations.
- Lattice approach: the information is represented at different levels of aggregation due to different levels of confidentiality.

These two model are however too restrictive and prone to attacks (see for example [1]) to be considered something more than a past history.

Cox and Sande [20, 84] propose a cell suppression technique. The idea is to suppress from the released database all the attributes considered confidential and the attributes that can be used to infer some confidential information (complementary suppression). Denning [25] shows that such an approach is unfeasible in case of complex queries. However, the identification of the set of cells to be suppressed is a not negligible problem especially in the case of complementary suppression. The problem was deeply investigated by Cox [20].

Recently, in the context of statistical disclosure control, a large number of methods, called *masking* methods in the SDC jargon, have been developed to preserve individual privacy when releasing aggregated statistics on data, and more specifically to anonymize the released statistics from those data items that can identify one among the individual entities (person, household, business, etc.) whose features are described by the statistics, also taking into account related information publicly available [106]. In [27] a description of the

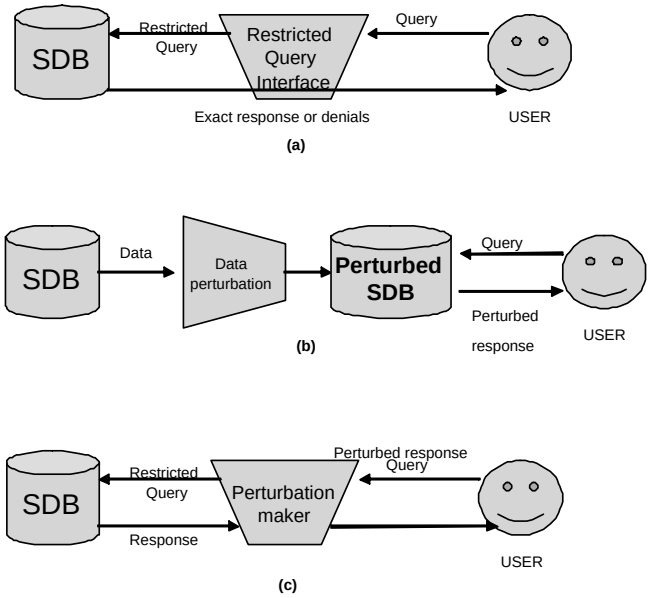

Figure 3.1: Three traditional strategies of SDC: (a) Query set restriction, (b) Data Perturbation, (c) Output Perturbation

most relevant masking methods proposed so far is presented. Among the perturbative methods specifically designed for continuous data, the following masking techniques are proposed: additive noise, data distortion by probability distribution, resampling, microaggregation, rank swapping. For categorical data both perturbative and non-perturbative methods are presented. The top-coding and bottom-coding techniques are both applied to ordinal categorical variables; they recode, respectively, the first/last p values of a variable into a new category. The global-recoding technique, instead, recodes the p lowest frequency categories into a single one. All these masking methods are assessed according to the two main parameters: the *information loss* and the *disclosure risk*, that is, the risk that a piece of information be linked to a specific individual. Several methods are presented in the paper for assessing the *information loss* and the *disclosure risk* given by a SDC method. Additionally, in order to provide a trade-off level between these two metrics, a score is defined that gives the same importance to disclosure risk and information loss.

Bibliography

[1] N. Adam and J. Worthmann, *Security-control methods for statistical databases: a comparative study.* ACM Comput. Surv., Volume 21(4), pp. 515-556, year 1989, ACM Press.

[2] D. Agrawal and C. C. Aggarwal, *On the Design and Quantification of Privacy Preserving Data Mining Algorithms.* In Proceedings of the 20*th* ACM Symposium on Principle of Database System, pp. 247-255, year 2001, ACM Press.

[3] R. Agrawal, T. Imielinski and A. Swami, *Mining Association Rules between Sets of Items in Large Databases.* Proceedings of ACM SIGMOD, pp. 207-216, May 1993, ACM Press.

[4] R. Agrawal and R. Srikant, *Privacy Preserving Data Mining.* In Proceedings of the ACM SIGMOD Conference of Management of Data, pp. 439-450, year 2000, ACM Press.

[5] R. Agrawal and R. Srikant, *Fast algorithms for mining association rules.* In Proceeding of the 20*th* International Conference on Very Large Databases, Santiago, Chile, June 1994, Morgan Kaufmann.

[6] G. M. AMDAHL, *Validity of the Single-Processor Approach to Achieving Large Scale Computing Capabilities.* AFIPS Conference Proceedings(April 1967),pp. 483-485, Morgan Kaufmann Publishers Inc.

[7] M. J. Atallah, E. Bertino, A. K. Elmagarmid, M. Ibrahim and V. S. Verykios, *Disclosure Limitation of Sensitive Rules.* In Proceedings of the IEEE Knolwedge and Data Engineering Workshop, pp. 45-52, year 1999, IEEE Computer Society.

[8] D. P. Ballou, H. L. Pazer, *Modelling Data and Process Quality in Multi Input, Multi Output Information Systems.* Management science, Vol. 31, Issue 2, pp. 150-162, (1985).

[9] Y. Bar-Hillel, *An examination of information theory.* Philosophy of Science, volume 22, pp.86-105, year 1955.

[10] E. Bertino, I. Nai Fovino and L. Parasiliti Provenza, *A Framework for Evaluating Privacy Preserving Data Mining Algorithms*. Data Mining and Knowledge Discovery Journal, year 2005, Kluwert.

[11] E.Bertino and I.Nai Fovino, *Information Driven Evaluation of Data Hiding Algorithms. 7th* International Conference on Data Warehousing and Knowledge Discovery. Copenaghen, August 2005, Springer-Verlag.

[12] N. M. Blachman, *The amount of information that y gives about X.* IEEE Truns. Inform. Theon. vol. IT-14, pp. 27-31. Jan. 1968, IEEE Press.

[13] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification of Regression Trees.* Wadsworth International Group, year 1984.

[14] S. Brin, R. Motwani, J. D. Ullman and S. Tsur, *Dynamic itemset counting and implication rules for market basket data.* In Proc. of the ACM SIGMOD International Conference on Management of Data, year 1997, ACM Press.

[15] L. Chang and I. S. Moskowitz, *Parsimonious downgrading and decision trees applied to the inference problem.* In Proceedings of the 1998 New Security Paradigms Workshop, pp.82-89, year 1998, ACM Press.

[16] P. Cheeseman and J. Stutz, *Bayesian Classification (AutoClass): Theory and Results.* Advances in Knowledge Discovery and Data Mining, AAAI Press/MIT Press, year 1996.

[17] M. S. Chen, J. Han and P. S. Yu, *Data Mining: An Overview from a Database Perspective.* IEEE Transactions on Knowledge and Data Engineering, vol. 8 (6), pp. 866-883, year 1996, IEEE Educational Activities Department.

[18] F. Y. Chin and G. Ozsoyoglu, *Auditing and inference control in statistical databases.* IEEE Trans. Softw. Eng. SE-8, 6 (Apr.), pp. 574-582, year 1982, IEEE Press.

[19] F. Y. Chin and G. Ozsoyoglu, *Statistical database design.* ACM Trans. Database Syst. 6, 1 (Mar.), pp. 113-139, year 1981, ACM Press.

[20] L. H. Cox, *Suppression methodology and statistical disclosure control.* J. Am. Stat. Assoc. 75, 370 (June), pp. 377-385, year 1980.

[21] E. Dasseni, V. S. Verykios, A. K. Elmagarmid and E. Bertino, *Hiding Association Rules by using Confidence and Support.* in proceedings of the *4th* Information Hiding Workshop, pp. 369383, year 2001, Springer-Verlag.

[22] D. Defays, *An efficient algorithm for a complete link method.* The Computer Journal, 20, pp. 364-366, 1977.

[23] D. E. Denning and J. Schlorer, *Inference control for statistical databases.* Computer 16 (7), pp. 69-82, year 1983 (July), IEEE Press.

[24] D. Denning, *Secure statistical databases with random sample queries.* ACM TODS, 5, 3, pp. 291-315, year 1980.

[25] D. E. Denning, *Cryptography and Data Security.* Addison-Wesley, Reading, Mass. 1982.

[26] V. Dhar, *Data Mining in finance: using counterfactuals to generate knowledge from organizational information systems.* Information Systems, Volume 23, Number 7, pp. 423-437(15), year 1998.

[27] J. Domingo-Ferrer and V. Torra, *A Quantitative Comparison of Disclosure Control Methods for Microdata.* Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 113-134, P. Doyle, J. Lane, J. Theeuwes, L. Zayatz ed., North-Holland, year 2002.

[28] P. Domingos and M. Pazzani, *Beyond independence: Conditions for the optimality of the simple Bayesian classifier.* Proceedings of the Thirteenth International Conference on Machine Learning, pp. 105-112, San Francisco, CA, year 1996, Morgan Kaufmann.

[29] P. Drucker, *Beyond the Information Revolution.* The Atlantic Monthly, 1999.

[30] P. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* Wiley, year 1973, New York.

[31] G. T. Duncan, S. A. Keller-McNulty and S. L. Stokes, *Disclosure risks vs. data utility: The R-U confidentiality map.* Tech. Rep. No. 121. National Institute of Statistical Sciences. 2001

[32] C. Dwork and K. Nissim, *Privacy preserving data mining in vertically partitioned database.* In Crypto 2004, Vol. 3152, pp. 528-544.

[33] D. L. EAGER, J. ZAHORJAN and E. D. LAZOWSKA, *Speedup Versus Efficiency in Parallel Systems.* IEEE Trans. on Computers, C-38, 3 (March 1989), pp. 408-423, IEEE Press.

[34] L. Ertoz, M. Steinbach and V. Kumar, *Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data.* In Proceeding to the SIAM International Conference on Data Mining, year 2003.

[35] M. Ester, H. P. Kriegel, J. Sander and X. XU, *A density-based algorithm for discovering clusters in large spatial databases with noise.* In Proceedings of the 2nd ACM SIGKDD, pp. 226-231, Portland, Oregon, year 1996, AAAI Press.

[36] A. Evfimievski, *Randomization in Privacy Preserving Data Mining.* SIGKDD Explor. Newsl., vol. 4, number 2, year 2002, pp. 43-48, ACM Press.

[37] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, *Privacy Preserving Mining of Association Rules.* In Proceedings of the 8*th* ACM SIGKDDD International Conference on Knowledge Discovery and Data Mining, year 2002, Elsevier Ltd.

[38] S. E. Fahlman and C. Lebiere, *The cascade-correlation learning architecture.* Advances in Neural Information Processing Systems 2, pp. 524-532. Morgan Kaufmann, year 1990.

[39] R. P. Feynman, R. B. Leighton and M. Sands, *The Feynman Lectures on Physics, v I.* Reading, Massachusetts: Addison-Wesley Publishing Company, year 1963.

[40] S. Fortune and J. Wyllie, *Parallelism in Random Access Machines.* Proc. Tenth ACM Symposium on Theory of Computing(1978), pp. 114-118, ACM Press.

[41] W. Frawley, G. Piatetsky-Shapiro and C. Matheus, *Knowledge Discovery in Databases: An Overview.* AI Magazine, pp. 213-228, year 1992.

[42] S. P. Ghosh, *An application of statistical databases in manufacturing testing.* IEEE Trans. Software Eng. 1985. SE-11, 7, pp. 591-596, IEEE press.

[43] S.P.Ghosh, *An application of statistical databases in manufacturing testing.* In Proceedings of IEEE COMPDEC Conference, pp. 96-103,year 1984, IEEE Press.

[44] S. Guha, R. Rastogi and K. Shim, *CURE: An efficient clustering algorithm for large databases.* In Proceedings of the ACM SIGMOD Conference, pp. 73-84, Seattle, WA. 1998, ACM Press.

[45] S. Guha, R. Rastogi and K. Shim, *ROCK: A robust clustering algorithm for categorical attributes.* In Proceedings of the 15*th* ICDE, pp. 512-521, Sydney, Australia, year 1999, IEEE Computer Society.

[46] J. Han and M. Kamber, *Data Mining: Concepts and Techniques.* The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, August 2000.

[47] J. Han, J. Pei and Y. Yin, *Mining frequent patterns without candidate generation.* In Proceeding of the 2000 ACM-SIGMOD International Conference on Management of Data, Dallas, Texas, USA, May 2000, ACM Press.

[48] M. A. Hanson and R. L. Brekke, *Workload management expert system - combining neural networks and rule-based programming in an operational application.* In Proceedings Instrument Society of America, pp. 1721-1726, year 1988.

[49] J. Hartigan and M. Wong, *Algorithm AS136: A k-means clustering algorithm.* Applied Statistics, 28, pp. 100-108, year 1979.

[50] A. Hinneburg and D. Keim, *An efficient approach to clustering large multimedia databases with noise.* In Proceedings of the 4*th* ACM SIGKDD, pp. 58-65, New York, year 1998, AAAI Press.

[51] T. Hsu, C. Liau and D.Wang, *A Logical Model for Privacy Protection.* Lecture Notes in Computer Science, Volume 2200, Jan 2001, pp. 110-124, Springer-Verlag.

[52] IBM Synthetic Data Generator. http://www.almaden.ibm.com/software/quest/Resources/datasets/syndata.html

[53] M. Kantarcioglu and C. Clifton, *Privacy Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data.* In Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 24-31, year 2002, IEEE Educational Activities Department.

[54] G. Karypis, E. Han and V. Kumar, *CHAMELEON: A hierarchical clustering algorithm using dynamic modeling.* COMPUTER, 32, pp. 68-75, year 1999.

[55] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley and Sons, New York, year 1990.

[56] W. Kent, *Data and reality.* North Holland, New York, year 1978.

[57] S. L. Lauritzen, *The em algorithm for graphical association models with missing data.* Computational Statistics and Data Analysis, 19 (2), pp. 191-201, year 1995, Elsevier Science Publishers B. V.

[58] W. Lee and S. Stolfo, *Data Mining Approaches for Intrusion Detection.* In Proceedings of the Seventh USENIX Security Symposium (SECURITY '98), San Antonio, TX, January 1998.

[59] A. V. Levitin and T. C. Redman, *Data as resource: properties, implications and prescriptions.* Sloan Management review, Cambridge, Vol. 40, Issue 1, pp. 89-101, year 1998.

[60] Y. Lindell and B. Pinkas, *Privacy Preserving Data Mining.* Journal of Cryptology, vol. 15, pp. 177-206, year 2002, Springer Verlag.

[61] R. M. Losee, *A Discipline Independent Definition of Information.* Journal of the American Society for Information Science 48 (3), pp. 254-269, year 1997.

[62] M. Masera, I. Nai Fovino, R. Sgnaolin *A Framework for the Security Assessment of Remote Control Applications of Critical Infrastructure* 29th ESReDA Seminar "Systems Analysis for a More Secure World", year 2005

[63] G. MClachlan and K. Basford, *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, New York, year 1988.

[64] M. Mehta, J. Rissanen and R. Agrawal, *MDL-based decision tree pruning.* In Proc. of KDD, year 1995, AAAI Press.

[65] G. L. Miller, *Resonance, Information, and the Primacy of Process: Ancient Light on Modern Information and Communication Theory and Technology.* PhD thesis, Library and Information Studies, Rutgers, New Brunswick, N.J., May 1987.

[66] I. S. Moskowitz and L. Chang, *A decision theoretical based system for information downgrading.* In Proceedings of the 5*th* Joint Conference on Information Sciences, year 2000, ACM Press.

[67] S. R. M. Oliveira and O. R. Zaiane, *Toward Standardization in Privacy Preserving Data Mining.* ACM SIGKDD 3rd Workshop on Data Mining Standards, pp. 7-17, year 2004, ACM Press.

[68] S. R. M. Oliveira and O. R. Zaiane, *Privacy Preserving Frequent Itemset Mining.* Proceedings of the IEEE international conference on Privacy, security and data mining, pp. 43-54, year 2002, Australian Computer Society, Inc.

[69] S. R. M. Oliveira and O. R. Zaiane, *Privacy Preserving Clustering by Data Transformation.* In Proceedings of the 18*th* Brazilian Symposium on Databases, Manaus, Amazonas, Brazil, pp. 304-318, year 2003.

[70] K. Orr, *Data Quality and System Theory.* Comm. of the ACM, Vol. 41, Issue 2, pp. 66-71, Feb. 1998, ACM Press.

[71] M. A. Palley and J. S. Simonoff, *The use of regression methodology for compromise of confidential information in statistical databases.* ACM Trans. Database Syst. 12,4 (Dec.), pp. 593-608, year 1987.

[72] J. S. Park, M. S. Chen and P. S. Yu, *An Effective Hash Based Algorithm for Mining Association Rules.* Proceedings of ACM SIGMOD, pp. 175-186, May, 1995, ACM Press.

[73] Z. Pawlak, *Rough SetsTheoretical Aspects of Reasoning about Data.* Kluwer Academic Publishers, 1991.

[74] G. Piatetsky-Shapiro, *Discovery, analysis, and presentation of strong rules.* Knowledge Discovery in Databases, pp. 229-238, AAAI/MIT Press, year 1991.

[75] A. D. Pratt, *The Information of the Image.* Ablex, Norwood, NJ, 1982.

[76] J. R. Quinlan, *C4.5: Programs for Machine Learning.* Morgan Kaufmann, year 1993.

[77] J. R. Quinlan, *Induction of decision trees.* Machine Learning, vol. 1, pp. 81-106, year 1986, Kluwer Academic Publishers.

[78] R. Rastogi and S. Kyuseok, *PUBLIC: A Decision Tree Classifier that Integrates Building and Pruning.* Data Mining and Knowledge Discovery, vol. 4, n.4, pp. 315-344, year 2000.

[79] R. Rastogi and K. Shim, *Mining Optimized Association Rules with Categorical and Numeric Attributes.* Proc. of International Conference on Data Engineering, pp. 503-512, year 1998.

[80] H. L. Resnikoff, *The Illusion of Reality.* Springer-Verlag, New York, 1989.

[81] S. J. Rizvi and J. R. Haritsa, *Maintaining Data Privacy in Association Rule Mining.* In Proceedings of the 28*th* International Conference on Very Large Databases, year 2003, Morgan Kaufmann.

[82] S. J. Russell, J. Binder, D. Koller and K. Kanazawa, *Local learning in probabilistic networks with hidden variables.* In International Joint Conference on Artificial Intelligence, pp. 1146-1152, year 1995, Morgan Kaufmann.

[83] D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Learning internal representations by error propagation.* Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations pp. 318–362, Cambridge, MA: MIT Press, year 1986.

[84] G. Sande, *Automated cell suppression to reserve confidentiality of business statistics.* In Proceedings of the 2nd International Workshop on Statistical Database Management, pp. 346-353, year 1983.

[85] A. Savasere, E. Omiecinski and S. Navathe, *An efficient algorithm for mining association rules in large databases.* In Proceeding of the Conference on Very Large Databases, Zurich, Switzerland, September 1995, Morgan Kaufmann.

[86] J. Schlorer, *Information loss in partitioned statistical databases.* Comput. J. 26, 3, pp. 218-223, year 1983, British Computer Society.

[87] C. E. Shannon, *A Mathematical Theory of Communication.* Bell System Technical Journal, vol. 27,(July and October),1948, pp.379423, pp. 623-656.

[88] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication.* University of Illinois Press, Urbana, Ill. 1949.

[89] A. Shoshani, *Statistical databases: characteristics,problems, and some solutions.* Proceedings of the Conference on Very Large Databases (VLDB), pp.208-222, year 1982, Morgan Kaufmann Publishers Inc.

[90] R. SIBSON, *SLINK: An optimally efficient algorithm for the single link cluster method.* Computer Journal, 16, pp. 30-34, year 1973.

[91] P. Smyth and R. M. Goodman, *An information theoretic Approach to Rule Induction from databases.* IEEE Transaction On Knowledge And Data Engineering, vol. 3, n.4, August,1992, pp. 301-316, IEEE Press.

[92] L. Sweeney, *Achieving* k*-Anonymity Privacy Protection using Generalization and Suppression.* International Jurnal on Uncertainty, Fuzzyness and Knowledge-based System, pp. 571-588, year 2002, World Scientific Publishing Co., Inc.

[93] R. Srikant and R. Agrawal, *Mining Generalized Association Rules.* Proceedings of the 21*th* International Conference on Very Large Data Bases, pp. 407-419, September 1995, Morgan Kaufmann.

[94] G. K. Tayi, D. P. Ballou, *Examining Data Quality.* Comm. of the ACM, Vol. 41, Issue 2, pp. 54-58, year 1998, ACM Press.

[95] M. Trottini, *A Decision-Theoretic Approach to data Disclosure Problems.* Research in Official Statistics, vol. 4, pp. 722, year 2001.

[96] M. Trottini, *Decision models for data disclosure limitation.* Carnegie Mellon University, Available at `http://www.niss.org/dgii/TR/ThesisTrottini -final.pdf`, year 2003.

[97] University of Milan - Computer Technology Institute - Sabanci University *Codmine* IST project. 2002-2003.

[98] J. Vaidya and C. Clifton, *Privacy Preserving Association Rule Mining in Vertically Partitioned Data.* In Proceedings of the 8*th* ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 639-644, year 2002, ACM Press.

[99] V. S. Verykios, E. Bertino, I. Nai Fovino, L. Parasiliti, Y. Saygin, Y. Theodoridis, *State-of-the-art in Privacy Preserving Data Mining.* SIGMOD Record, 33(1) pp. 50-57, year 2004, ACM Press.

[100] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin and E. Dasseni, *Association Rule Hiding.* IEEE Transactions on Knowledge and Data Engineering, year 2003, IEEE Educational Activities Department.

[101] C. Wallace and D. Dowe, *Intrinsic classification by MML. The Snob program.* In the Proceedings of the 7*th* Australian Joint Conference on Artificial Intelligence, pp. 37- 44, UNE, World Scientific Publishing Co., Armidale, Australia, 1994.

[102] G. J. Walters, *Philosophical Dimensions of Privacy: An Anthology.* Cambridge University Press, year 1984.

[103] G. J. Walters, *Human Rights in an Information Age: A Philosophical Analysis.* chapter 5, University of Toronto Press, year 2001.

[104] Y. Wand and R. Y. Wang, *Anchoring Data Quality Dimensions in Ontological Foundations.* Comm. of the ACM, Vol. 39, Issue 11, pp. 86-95, Nov. 1996, ACM Press.

[105] R. Y. Wang and D. M. Strong, *Beyond Accuracy: what Data Quality Means to Data Consumers.* Journal of Management Information Systems Vol. 12, Issue 4, pp. 5-34, year 1996.

[106] L. Willenborg and T. De Waal, *Elements of statistical disclosure control.* Lecture Notes in Statistics Vol.155, Springer Verlag, New York.

[107] N. Ye and X. Li, *A Scalable Clustering Technique for Intrusion Signature Recognition.* 2001 IEEE Man Systems and Cybernetics Information Assurance Workshop, West Point, NY, June 5-6, year 2001, IEEE Press.

[108] M. J. Zaki, S. Parthasarathy, M. Ogihara and W. Li, *New algorithms for fast discovery of association rules* In Proceeding of the 8rd International Conference on KDD and Data Mining, Newport Beach, California, August 1997, AAAI Press.

European Commission

EUR 23068 EN – Joint Research Centre – Institute for the Protection and Security of the Citizen
Title: State of the Art in Privacy Preserving Data Mining
Author(s): Igor Nai Fovino and Marcelo Masera
Luxembourg: Office for Official Publications of the European Communities
2008 – 51 pp. – 21 x 29.7 cm
EUR – Scientific and Technical Research series – ISSN 1018-5593

Abstract
Privacy is one of the most important properties an information system must satisfy. A relatively new trend shows that classical access control techniques are not sufficient to guarantee privacy when *Data Mining* techniques are used. Such a trend, especially in the context of public databases, or in the context of sensible information related to critical infrastructures, represents, nowadays a not negligible thread.
Privacy Preserving Data Mining (PPDM) algorithms have been recently introduced with the aim of modifying the database in such a way to prevent the discovery of sensible information.
This is a very complex task and there exist in the scientific literature some different approaches to the problem. In this work we present a ``Survey'' of the current PPDM methodologies which seem promising for the future

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.